Autor: Marcelo Cesário Miguel (marcelocm8@al.insper.edu.br)
Repositório: https://github.com/MarceloCMiguel/nlp-rating-analysis

Análise de Comentários e Avaliações de Produtos com Processamento de Linguagem Natural (NLP)

1. Introdução

A análise de comentários e avaliações de produtos é uma tarefa importante em muitas empresas, pois ajuda a entender o que os clientes estão dizendo sobre seus produtos e serviços. Nesse contexto, o uso de técnicas de processamento de linguagem natural (NLP) é essencial para extrair insights significativos desses dados.

No problema apresentado, foi fornecido um conjunto de dados contendo 3 principais colunas: título do comentário, comentário e a avaliação do produto, classificado em um sistema de estrelas de 0 a 5. A premissa assumida é que avaliações com 3 estrelas ou menos são negativas, enquanto avaliações com 4 ou 5 estrelas são positivas. A partir dessas informações, é possível entender melhor o que os clientes estão dizendo e identificar padrões nas opiniões dos consumidores.

2. Metodologia

2.1. Pré Processamento

A metodologia utilizada para a análise dos dados começou com a remoção de linhas vazias baseadas nas colunas de comentário e avaliação (estrelas de 0 a 5), a fim de evitar realizar manipulação de palavras em textos vazios ou em formato *pandas.na*.

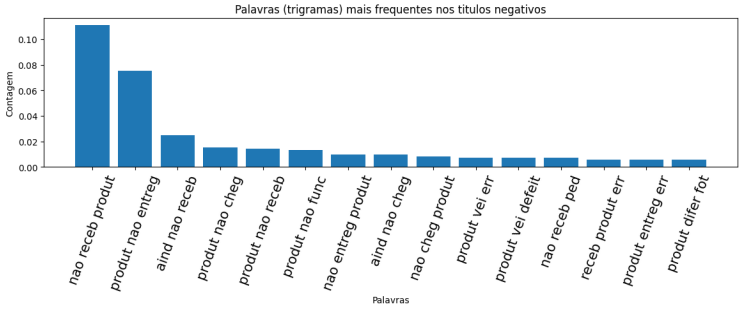
Em seguida, foi aplicado um processo de pré-processamento nos dados de texto (coluna de título e de comentário) para remover informações desnecessárias e padronizar o formato das palavras. Para isso, foi utilizada uma função que, em primeiro lugar, converte o texto para minúsculas, remove acentos, pontuações e números do texto.

Por fim, é aplicado um processo de remoção de palavras irrelevantes (*stop words*) seguido de *stemming*, onde cada palavra é reduzida ao seu radical, para que a análise considere apenas as raízes das palavras e não todas as suas variações, reduzindo a variabilidade do vocabulário e transformando palavras com variações flexionais e derivacionais em sua forma raiz, o que pode melhorar a eficácia de modelos de análise de texto.

2.2. Modelo

Para analisar esses dados, utilizou-se o CountVectorizer, uma técnica de vetorização de texto comum em NLP. Além disso, foi utilizado o argumento de trigramas no processo, o que permite capturar mais informações do contexto da frase, em vez de considerar apenas palavras isoladas como 'ruim' ou 'horrível'.

Após isso, tem-se o seguinte gráfico de frequência de palavras para títulos negativos:



Sendo assim, alguns dos fatores que mais influenciam em avaliações negativas, a partir do radical, são: não recebo produto, produto não entregue, produto não funciona.

Além disso, é possível analisar apenas os adjetivos utilizados durante as reclamações, utilizando unigramas na análise. Para isso, é necessário remover as palavras que aparecem na lista de palavras positivas. Essa abordagem ajuda a se concentrar nos termos mais relevantes para a análise de sentimento.

Após isso, tem-se o seguinte gráfico de frequência de palavras negativas filtradas: